A general framework for combining ecosystem models

Michael A. Spence[1,2,3*], Julia L. Blanchard[4], Axel G. Rossberg[3,5], Michael R. Heath[6], Johanna J Heymans[7], Steven Mackinson[3,8], Natalia Serpetti[7], Douglas C. Speirs[6], Robert B. Thorpe[3] and Paul G. Blackwell[1]

[1]School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

[2]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

[3]Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, Suffolk NR33 0HT, UK

[4]Institute for Marine and Antarctic Studies and Centre for Marine Socioecology, University of Tasmania, 20 Castray Esplanade, Battery Point. TAS. 7004

[5]Aquatic Ecology Group, Department of Organismal Biology, School of Biological and Chemical Sciences, Queen Mary University of London,

Mile End Road, London E1 4NS

[6]Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, Scotland

[7]Scottish Association for Marine Science, Scottish Marine Institute, Oban, Argyll, PA371QA

[8]Scottish Pelagic Fishermen's Association, Heritage House, 135 - 139 Shore Street, Fraserburgh, Aberdeenshire, AB43 9BP

[*]Corresponding author: michael.spence@cefas.co.uk

Running title: Combining ecosystem models

Abstract

When making predictions about ecosystems, we often have available a number of different ecosystem models that attempt to represent their dynamics in a detailed mechanistic way. Each of these can be used as a simulator of large-scale experiments and make projections about the fate of ecosystems under different scenarios in order to support the development of appropriate management strategies. However, structural differences, systematic discrepancies and uncertainties lead to different models giving different predictions. This is further complicated by the fact that the models may not be run with the same functional groups, spatial structure or time scale. Rather than simply trying to select a 'best' model, or taking some weighted average, it is important to exploit the strengths of each of the models, while learning from the differences between them. To achieve this, we construct a flexible statistical model of the relationships between a collection of mechanistic models

and their biases, allowing for structural and parameter uncertainty and for different ways of representing reality. Using this statistical meta-model, we can combine prior beliefs, model estimates and direct observations using Bayesian methods, and make coherent predictions of future outcomes under different scenarios with robust measures of uncertainty. In this paper we take a diverse ensemble of existing North Sea ecosystem models and demonstrate the utility of our framework by applying it to answer the question what would have happened to demersal fish if fishing was to stop.

Key-words: Bayesian statistics, Complex models, Multi-model ensemble, Multi-species models, Simulation models, Uncertainty analysis

1 Introduction

Ecosystem models are widely used to support policy decisions, including fisheries and marine environmental policies (Hyder et al 2015). Any such model is imperfect, and in order to use it to inform policy making, it is important to quantify the uncertainty of its predictions in a robust manner (Harwood and Stokes 2003; Williams and Hooten 2016). Often several models are available, each embodying some knowledge of a given ecosystem, but differing in their predictions. Choosing to use one model's prediction whilst excluding the others is limiting the amount of information available and therefore increasing uncertainty. Our aim here is to describe and demonstrate a framework for combining information from multiple ecosystem models in a coherent way that, following Chandler (2013), exploits their strengths and discounts their weaknesses.

Many methods of combining outputs from different models have been previously proposed. One is to use a 'democracy' of simulators (Payne et al 2015; Knutti 2010), where each model gets one vote, regardless of how well it represents the true system, and a distribution of possible outputs comes from this. Similarly, one could take an average of the model outputs, which often outperforms all the individual models (Rougier 2016). However, some models are better at predicting some outputs than others. An alternative approach is to try and find the 'best' model(s) (Payne et al 2015; Johnson and Omland 2004). These methods imply that at least one of the models is 'correct', in the sense that it can predict the true output. Not only is

this a bold assumption, but the addition of another model may allow an area of the output space to become probable when before it was not. Thus, by increasing the number of models there is no guarantee that the uncertainty will reduce. One way of deciding which model is the 'best' is to weight models using Bayes factors, also known as Bayesian model averaging (Banner and Higgs 2017; Ianelli et al 2016). As Chandler (2013) explains, there is generally no model better in all respects than the others and so there is no natural way of assigning a single weight to each model. Furthermore, if model outputs are not presented with uncertainty then, in the case where the truth is a continuous quantity, a simulator will almost never be 'correct', thus the probability of getting the true value from the ensemble is zero. Recently, 'ensemble models' have been used to describe how model outputs related to reality (Anderson et al 2017).

Applying the above methods to ecosystem models is not straightforward, as different models have often been fitted to different data (Ianelli et al 2016), and often their outputs are on different scales or represent different dynamical processes, which are sometimes integrated out. A further difficulty in applying these methods is that the ecosystem models can have different outputs that are not directly comparable. For example, whole ecosystem models often reduce complexity through the use of functional groups (e.g. Heath 2012) whereas partial ecosystem or multi-species models may focus on a reduced number of species (e.g. Blanchard et al 2014). However, different ecosystem models are often developed with similar underlying theory (e.g. food web interactions), could have similar dynamics and may even be developed in the same research groups (e.g. Heath (2012) and Speirs et al (2010)). They may also have similar forcing inputs, for example those coming from global regional physical or biogeochemical models such as those used in model inter-comparison studies (e.g. Tittensor et al 2017). When combining model

outputs, it is important to take these similarities into account rather than treating the models as independent (Rougier et al 2013).

Another approach is to think of the ecosystem models as coming from a population of such models (Tebaldi and Sansó 2009; Chandler 2013; Leith and Chandler 2010) and then describe how the population differs from reality. It makes sense that several models in an ensemble model would inform one another. For example, one model (m1) may contain several demersal fish species and the other (m2) a functional group called "demersal fish". Although m2 does not explicitly contain the species Atlantic cod (*Gadus morhua*) its relationship with m1 may be able to tell us something about Atlantic cod indirectly. In other words, modelling the models allows us to sample the unobserved outputs, conditional on the models' observed outputs.

In this paper we describe an ensemble model which is based on the principles of Chandler (2013) but which models the outputs themselves, varying in form between the different ecosystem models, rather than statistical descriptors of the outputs. Our approach involves statistical modelling of the relationship between an 'ensemble' of ecosystem models. To avoid ambiguity, we will refer to the latter henceforth as 'simulators' and we refer to the way in which a simulator output differs from reality as its discrepancy. As we are interested in measuring uncertainty our statistical modelling will apply Bayesian inference methods (Robert 2007), and our analysis will consider any relevant prior knowledge as well as simulator outputs that predict what would happen in the future under different management scenarios. The Bayesian approach is subjective; for an introduction to subjective uncertainty and decision theory, see Berger (1985). Strictly speaking, any fully Bayesian analysis involves obtaining the posterior beliefs of a particular individual, by combining their prior beliefs with information from data and modelling. Depending on the

context, that individual may be, for example, either a scientist or a policy maker. Our framework includes the elicitation of prior beliefs to combine with information from the model ensemble, allowing different individuals' posterior distributions to be obtained. For the purpose of our case study, the individual chosen is one of the authors.

In Section 2 we set up the general framework and in Section 3 we demonstrate the model by looking at a specific case study: what would have happened in the North Sea if we had stopped fishing in 2014? We conclude by discussing wider applications of the approach in Section 4.

2 General framework

We think of the available simulators as coming from some conceptual population. Our *a priori* beliefs about each one are the same; we are treating the simulators as unlabelled 'black boxes'. More formally, we regard the simulators as 'exchangeable'; see Gelman et al (2013). We consider relaxing this assumption in Section 4. This idea is formalised by using a hierarchical model (for more information see Gelman et al (2013)) to represent the ensemble of simulators. However, there is no reason to believe that the population of simulators will either contain, or be centred on, the truth (Chandler 2013) so we need to allow some difference between the population of simulators and the truth.

To describe the relationship between the simulators and the truth we developed an ensemble model that describes the population of simulators, its dynamics and its relation with the true quantity of interest. We are interested in n true quantities, $\boldsymbol{y}^{(t)} = (y_1^{(t)}, \ldots, y_n^{(t)})'$, e.g. the biomass of n species at a time t, for times $t = 1, \ldots, T$. We regard m simulators, each giving an output representing the quantities

of interest, $\boldsymbol{x}_i^{(t)} = (x_{i1}^{(t)}, \ldots, x_{in}^{(t)})'$ for $i = 1, \ldots, m$, as coming from a population with expected output $\boldsymbol{\mu}^{(t)} = (\mu_1^{(t)}, \ldots, \mu_n^{(t)})'$, the simulator consensus. To define our ensemble model, we describe separately the difference between $\boldsymbol{y}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$, the shared discrepancy, and the difference between $\boldsymbol{x}_i^{(t)}$ and $\boldsymbol{\mu}^{(t)}$, simulator i's individual discrepancy. Figure 1 illustrates an example of the ensemble model at time t. It can be read as a geometrical representation of how the simulators and reality relate to one another (see also Chandler 2013). In the subsequent subsections we describe the specific details of the general ensemble model. A summary of the variables and the model can be found in Table 1.

2.1 Uncertainty in simulator outputs

The outputs from simulator i, an n_i dimensional vector $\boldsymbol{u}_i^{(t)}$, may not always represent the elements of $\boldsymbol{x}_i^{(t)}$, its 'best guess', directly. For example, the elements of $\boldsymbol{x}_i^{(t)}$ may represent biomasses of individual fish species and the elements of $\boldsymbol{u}_i^{(t)}$ may represent the biomass of functional groups, e.g. biomass of demersal fish.

We say that

$$\boldsymbol{u}_i^{(t)} = f_i(\boldsymbol{x}_i^{(t)}),$$

for some simulator-specific function $f_i(\cdot)$. For example, if the elements of $\boldsymbol{u}_i^{(t)}$ are elements of $\boldsymbol{x}_i^{(t)}$ or are sums of those elements, perhaps with some rescaling, then the relationship is linear

$$\boldsymbol{u}_i^{(t)} = M_i \boldsymbol{x}_i^{(t)},$$

where M_i is an $n_i \times n$ matrix. For other examples see Table 2.

Generally the simulators are run with uncertain inputs and parameter values. This leads to uncertainty in the outputs and is commonly known as parameter

uncertainty. We say that

$$\boldsymbol{u}_i^{(t)} = \hat{\boldsymbol{u}}_i^{(t)} + \boldsymbol{\epsilon}_{u_i},$$

for $t \in S_i$, where $\boldsymbol{\epsilon}_{u_i}$ has expectation $\mathbf{0}$ and is sampled from a simulator-specific distribution and $\hat{\boldsymbol{u}}_i^{(t)}$ is the expectation of the ith simulator's output at time t. The simulator-specific distribution is found from fitting the simulator to a finite dataset (e.g. Spence et al 2016; Thorpe et al 2015) or by performing sensitivity analysis of the simulator inputs (e.g. Morris et al 2014).

2.2 Individual discrepancy

At time t, the difference between simulator i's 'best guess', $\boldsymbol{x}_i^{(t)}$, and the simulator consensus, $\boldsymbol{\mu}^{(t)}$, is simulator i's individual discrepancy,

$$\boldsymbol{x}_i^{(t)} - \boldsymbol{\mu}^{(t)} = \boldsymbol{\gamma}_i + \boldsymbol{z}_i^{(t)}.$$

This divides the individual discrepancy between the long-term individual discrepancy, $\boldsymbol{\gamma}_i$, and the short-term individual discrepancy, $\boldsymbol{z}_i^{(t)}$. $\boldsymbol{\gamma}_i$ is an n dimensional random variable with expectation $\mathbf{0}$ and covariance C. It seems natural to allow $\boldsymbol{z}_i^{(t)}$ and $\boldsymbol{z}_i^{(t+1)}$ to be dependent on each other; for example, if at time t, $\boldsymbol{z}_i^{(t)}$ was less than $\mathbf{0}$, then we might also expect $\boldsymbol{z}_i^{(t+1)}$ to be less than $\mathbf{0}$. With this in mind, we say that $\boldsymbol{z}_i^{(t)}$ follows a stationary auto-regressive model of order 1,

$$\boldsymbol{z}_i^{(t)} = R_i \boldsymbol{z}_i^{(t-1)} + \boldsymbol{\epsilon}_{z,t,i}, \tag{1}$$

for $t > 1$, where each $\boldsymbol{\epsilon}_{z,t,i}$ is an independent n-dimensional random variable centred on $\mathbf{0}$ with covariance Λ_i and R_i is an $n \times n$ matrix with the constraint such that R_i is stable, i.e. $\lim_{k\to\infty} R_i^k = 0$. R_i and Λ_i describe the dynamics of simulator i with $R_i \sim g_R(\cdot)$ and $\Lambda_i \sim g_\Lambda(\cdot)$ for some distributions g_R and g_Λ. At $t = 1$, $\boldsymbol{z}_i^{(1)}$ is sampled from the stationary distribution of the auto-regressive model described

in equation 1 (See Appendix A for more details). This formulation means that the expectation of the long-run behaviour of the individual discrepancy is the long-term individual discrepancy, i.e.

$$\begin{aligned}\lim_{k\to\infty} E(\boldsymbol{\gamma}_i + \boldsymbol{z}_i^{(t+k)} | \boldsymbol{\gamma}_i + \boldsymbol{z}_i^{(t)}) &= \boldsymbol{\gamma}_i + \lim_{k\to\infty} E(\boldsymbol{z}_i^{(t+k)} | \boldsymbol{z}_i^{(t)}) \\ &= \boldsymbol{\gamma}_i + E(\boldsymbol{z}_i^{(t)}) \\ &= \boldsymbol{\gamma}_i.\end{aligned}$$

2.3 Shared discrepancy

The shared discrepancy, the difference between the simulator consensus, $\boldsymbol{\mu}^{(t)}$, and truth, $\boldsymbol{y}^{(t)}$, is split up into the long-term shared discrepancy, $\boldsymbol{\delta}$, and the short-term shared discrepancy, $\boldsymbol{\eta}^{(t)}$, i.e.

$$\boldsymbol{y}^{(t)} - \boldsymbol{\mu}^{(t)} = \boldsymbol{\delta} + \boldsymbol{\eta}^{(t)}.$$

The short-term shared discrepancy is described by a stationary auto-regressive model of order 1

$$\boldsymbol{\eta}^{(t)} = R_\eta \boldsymbol{\eta}^{(t-1)} + \boldsymbol{\epsilon}_{\eta,t}, \tag{2}$$

for $t > 1$, where R_η is stable and $\boldsymbol{\epsilon}_{\eta,t}$ is an n dimensional random variable centred on $\mathbf{0}$ with covariance Δ. At $t = 1$, $\boldsymbol{\eta}^{(1)}$ is sampled from the stationary distribution of the auto-regressive model described in equation 2 (See Appendix A for more details). This formulation means that the expectation of the long-run behaviour of the shared discrepancy is the long-term shared discrepancy, i.e.

$$\begin{aligned}\lim_{k\to\infty} E(\boldsymbol{\delta} + \boldsymbol{\eta}^{(t+k)} | \boldsymbol{\delta} + \boldsymbol{\eta}^{(t)}) &= \boldsymbol{\delta} + \lim_{k\to\infty} E(\boldsymbol{\eta}^{(t+k)} | \boldsymbol{\eta}^{(t)}) \\ &= \boldsymbol{\delta} + E(\boldsymbol{\eta}^{(t)}) \\ &= \boldsymbol{\delta}.\end{aligned}$$

2.4 The truth

In the absence of any simulators, our prior beliefs for the truth at time t, $\boldsymbol{y}^{(t)}$, follow a random walk,

$$\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)} + \boldsymbol{\epsilon}_{\Lambda,t},$$

for $t > 1$, where each $\boldsymbol{\epsilon}_{\Lambda,t}$ is centred on $\mathbf{0}$ with covariance Λ_y. At $t = 1$, the truth, $\boldsymbol{y}^{(1)}$, follows a generic prior distribution $p(\boldsymbol{y}^{(1)})$.

At times $t \in S_0$, there are n_y noisy and possibly indirect observations, $\hat{\boldsymbol{w}}^{(t)}$, of the truth which come from some distribution, $p(\hat{\boldsymbol{w}}^{(t)}|\boldsymbol{y}^{(t)})$ that is problem specific and is caused by data uncertainty (Li and Wu 2006). The elements of $\hat{\boldsymbol{w}}^{(t)}$ may not be the same as that of $\boldsymbol{y}^{(t)}$, for example if observations are incomplete or aggregated. We assume that the sampling distribution of observations depends on the truth through some function $f_y(\cdot)$, such that

$$\boldsymbol{w}^{(t)} = f_y(\boldsymbol{y}^{(t)})$$

and $p(\hat{\boldsymbol{w}}^{(t)}|\boldsymbol{y}^{(t)}) = p(\hat{\boldsymbol{w}}^{(t)}|\boldsymbol{w}^{(t)})$.

For example if $\boldsymbol{w}^{(t)}$ is some linear transformation of $\boldsymbol{y}^{(t)}$, then

$$\boldsymbol{w}^{(t)} = M_y \boldsymbol{y}^{(t)}$$

where $M_{\boldsymbol{y}}$ is an $n_y \times n$ matrix.

3 Case Study

We illustrate our model by looking at a problem where a scientist needs to formally summarise uncertain model results, for example to present to other scientists or to decision makers about what would happen to the biomass of demersal species in the North Sea if fishing were to stop completely in 2014. We use outputs from five ecosystem simulators: Ecopath with Ecosim (EwE, Lynam and Mackinson 2015),

mizer (Blanchard et al 2014), FishSUMs (Speirs et al 2010), StrathE2E (Heath et al 2014) and LeMans (Thorpe et al 2015) (see Appendix B for more details about the simulators), as well as data from the International Bottom Trawl Survey (IBTS) (ICES Database of Trawl Surveys (DATRAS) 2015). In this example, one of the authors, JLB, has taken this role. Her prior beliefs are elicited and expressed as a prior distribution and the posterior distribution captures her uncertainty about the future of the ecosystem in this scenario give the relationships among the simulators and observations.

3.1 Groups of species

The five simulators represent demersal fish in different ways, with different species resolution and coverage. While our main interest is in demersal fish collectively, we need to represent the state of the ecosystem at a resolution that enables us to link these simulator outputs together.

In representing the state of the ecosystem, it would be computationally inefficient to treat each species separately, given that we are interested in demersal fish in aggregate. Instead, we can reduce the dimension of the problem by grouping the species together. This grouping needs to have the property that any simulator output that we can use can be expressed as the sum of one or more of our groups. The groups do not necessarily need to have any direct biological interpretation; provided the groups meet the criterion above, and allow us to represent the quantities of interest—here, demersal fish, given by the sum of all groups—the precise choice will not affect the answer obtained. For computational efficiency, we choose the minimum number of groups that meets this criterion while covering all demersal species. For example we grouped together monkfish, long rough dab, lemon sole and witch because they all occur in exactly the same simulators, as individual

species in EwE and LeMans and implicitly in StrathE2E, but are not contained in any larger set of species for which this is true. This minimal set consists of 5 groups, which we will model explicitly. The groups are:

1. *Common demersal*: These are Atlantic cod (*Gadus morhua*), haddock (*Melanogrammus aeglefinus*), whiting (*Merlangius merlangus*), Norway pout (*Trisopterus esmarkii*), European plaice (*Pleuronectes platessa*), common dab (*Limanda limanda*) and grey gurnard (*Eutrigla gurnardus*).
2. *Sole*: This is common sole (*Solea solea*).
3. *Monkfish etc.*: These are monkfish (*Lophius piscatorius*), long rough dab (*Hippoglossoides platessoides*), lemon sole (*Microstomus kitt*) and witch (*Glyptocephalus cynoglossus*).
4. *Poor Cod and Rays*: These are poor cod (*Trisopterus minutus*), starry rays (*Amblyraja radiata*) and cuckoo rays (*Leucoraja naevus*).
5. *Other demersal fish*: This consists of all other demersal fish.

We consider the total biomass densities for each of these groups, in tonnes per square kilometre, modelled on the log scale (to base 10, for ease of interpretation).

3.2 Data and elements of the statistical model

The IBTS data were extracted as in Fung et al (2012), to reveal the total catch on the survey for each of the 5 groups for the first (1986-2013) and third quarter (1991-2013). How this value relates to the true biomass density in the North Sea is not trivial, and these values are often multiplied by catchability coefficients (Walker et al 2017) which are themselves uncertain and model-based. In this example we are only interested in the biomass density relative to 2010, and therefore the total catch from the IBTS survey is enough provided we assume that catchability coefficients

are constant over time. Thus each element of $\boldsymbol{y}_t$ represents the log to base 10 of the total biomass (tonnes per kilometre squared) for one of our groups of species, averaged over year t, relative to 2010. Therefore,

$$\boldsymbol{w}^{(t)} = f_y(\boldsymbol{y}^{(t)}) = 10^{\boldsymbol{y}^{(t)}}.$$

The measurement error on the observations of the truth is assumed to be normally distributed on the $\log_{10}$ scale such that

$$\log_{10}\left(\hat{\boldsymbol{w}}^{(t)}/\hat{\boldsymbol{w}}^{(2010)}\right) \sim \mathrm{N}(\boldsymbol{y}^{(t)}, \Sigma_y),$$

for $t \neq 2010$. In this work we take Σ_y to be $2\log_{10}(1.15)$ on the diagonal elements and 0 on the off diagonal elements. This was chosen so that it means that the standard deviation of the true biomass would be 15% of the actual amount caught.

3.3 Simulators

We have outputs from five different simulators all of which have been run with zero fishing pressure from 2014 onwards. A short summary of the simulators, their outputs with respect to this case study and their simulator-specific function, $f_i(\cdot)$, can be found in Table 2. The ith simulator's output is assumed to be normally distributed on the $\log_{10}$ scale,

$$\log_{10} \boldsymbol{u}_i^{(t)} \sim \mathrm{N}(\log_{10} \hat{\boldsymbol{u}}_i^{(t)}, \Sigma_i),$$

with Σ_i fitted based on running simulator i many times (Leith and Chandler 2010; Chandler 2013). However, if this was not the case Σ_i could be estimated within the hierarchical system.

3.4 Ensemble model

Each element of $\boldsymbol{x}_i^{(t)}$ is the "best guess" of simulator i of the elements of $\boldsymbol{y}^{(t)}$, for $t = 1968, \ldots, 2100$, in log (base 10) tonnes per km squared of wet biomass. In this example we expect each of the simulators to converge to its own steady state, given that all external drivers are constant. This means that in equation 1 we expect R_i to tend towards 1 and Λ_i to tend towards 0. Furthermore, if a simulator reaches a stationary state before it has stopped running, then we know that it will be in that state forever. Simulator i's individual discrepancy, $\boldsymbol{\gamma}_i + \boldsymbol{z}_i^{(t)}$, is thus modelled as

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(0, C)$$

and

$$\boldsymbol{z}_i^{(t)} \sim \begin{cases} \mathrm{N}(R_i \boldsymbol{z}_i^{(t-1)}, \Lambda_i) & \text{if } t \leq 2013, \\ \mathrm{N}(h_z(R_i, k_i, t)\boldsymbol{z}_i^{t-1}, h_\Lambda(t, k_i)\Lambda_i) & \text{if } 2014 \geq t. \end{cases}$$

where

$$h_z(R_i, k, t) = R_i + (1 - R_i)(1 - h_\Lambda(t, k_i))$$

and

$$h_\Lambda(t, k_i) = \exp\left\{-k_i\,(t - 2013)\right\}.$$

This is saying that, after the end of fishing, the variance of the truth of model i reduces and the amount that the last value of $\boldsymbol{z}_i^{(t)}$ relates to the next moves towards 1 by a factor of $\exp(k_i)$ each year. We take $k_i \in [0, 6]$, as there is not much difference numerically if k_i goes above 6, with

$$k_i/6 \sim \mathrm{Beta}(a_k, b_k).$$

The diagonal elements of R_i fall between -1 and 1 with

$$\frac{R_i + 1}{2} \sim \mathrm{Beta}(\boldsymbol{a}_R, \boldsymbol{b}_R)$$

and the off-diagonal elements are set to 0. The simulator-specific variance parameter, Λ_i, is decomposed into a diagonal matrix of variances, Π_i, and a correlation matrix, P_i, such that

$$\Lambda_i = \Pi_i P_i \Pi_i. \tag{3}$$

The form of the prior distribution for the jth diagonal element of Π_i was

$$\pi_{ij} \sim \text{Gamma}(\alpha_{\pi,j}, \beta_{\pi,j}).$$

Distributions over correlation matrices are complicated by the mathematical requirement of positive definiteness. In practice, we specify separate priors on the elements, and then condition on positive definiteness; the unconditional prior for the j,kth element of P_i is given by

$$\frac{\rho_{ijk}+1}{2} \sim \begin{cases} \text{Beta}(a_{\rho jk}, b_{\rho jk}) & \text{if } j \neq k, \\ 1 & \text{otherwise.} \end{cases}$$

The difference between the truth at time t and the corresponding simulator consensus, $\boldsymbol{\mu}^{(t)}$, is then

$$\left(\boldsymbol{y}^{(t)}\right) - \left(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(2010)}\right) = \boldsymbol{\eta}^{(t)} + \boldsymbol{\delta}$$

with

$$\boldsymbol{\eta}^{(t)} \sim \text{N}(R_\eta \boldsymbol{\eta}^{(t-1)}, \Delta_\eta). \tag{4}$$

When the fishing is turned off, we are particularly uncertain about what will happen; thus we will remove any direct relation between $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t+1}$ beyond that time. We will say that

$$\boldsymbol{\mu}^{(t)} \sim \text{N}(\boldsymbol{\mu}^{(t-1)}, h_\Lambda(t, k_\mu)\Delta_\mu) \tag{5}$$

where $k_\mu \in [0, 6]$, so that the simulator consensus reaches a stationary point, as the individual simulators do.

We focus on the subjective probabilities of a particular individual, in this case JLB. Her prior beliefs were elicited using the method described in O'Hagan et al (2006) and Alhussain and Oakley (2017). Details of the prior elicitation can be found in Appendix C. Due to the dimensionality and correlation of the uncertain parameter space, we fitted the model using No U-turn Hamiltonian Monte Carlo (Hoffman and Gelman 2014) in the package Stan (Gelman et al 2015).

3.5 Results

The ensemble model predictions show changes in the uncertainty of relative biomass over time for each group of species, including projections following a fishing closure in 2014 (Figure 2). Each plot shows the marginal posterior distributions of each element of $\boldsymbol{y}^{(t)}$, for all t. Unsurprisingly, the ensemble model predicts *common demersal* fish increase following the fishery closure, as this group contains a lot of species targeted by fisheries.

According to the ensemble model the probability that there will be a greater total biomass of *common demersal* in 2050 than in 2010 is 0.90. There is a similar number for *sole* (0.93) and for *monkfish etc.* (0.88) but it is lower for *poor cod and rays* (0.55) and for the *other demersal* species (0.17).

The ensemble model also 'predicts' what happened before the data; that is, it gives posterior distributions for the actual values given the imperfect data and the simulator runs. Only *sole* and *common demersal* are output by simulators prior to 1986 and this is reflected in the increased uncertainty as we move further back in time from 1986.

The uncertainty in the prediction increases the further away from the observations of the truth, both when projecting and hindcasting. The uncertainty also increases when there are fewer simulators that give outputs. All of the simulators

give outputs for the *common demersal* group, four explicitly and one implicitly, and therefore we are more certain about what will happen to it in the future than for *poor cod and rays*, where only three simulators predict values for the future and only one explicitly. The uncertainty is highest for *other demersal* species. This is understandable as only two simulators predict values for this group of species, neither of which does so explicitly.

The absolute total biomass of demersal species is difficult to calculate here without information on the discrepancy between the simulator consensus and the truth. Although survey data are available, their relationship with the truth depends on the varying, and unknown, catchability coefficients for each of the groups. Although catchabilities can be estimated, for simplicity here we examine the total demersal biomass under the assumption that the groups had the same catchability coefficients (Figure 3). Again there is high uncertainty about whether the biomass will grow relative to the biomass in 2010. However, what it was before 1986 is also quite uncertain. This is because of the uncertainty in the populations of *Other demersal species*.

The median "best guess" of each of the simulators can also be compared across the different simulators (Figure 4). StrathE2E predicts quite a large increase in *common demersal* despite not explicitly outputting it. Mizer does not do a very good job of predicting the dynamics of *sole*, therefore the dynamics of the simulator consensus do not match the dynamics of mizer.

The posterior predictive distribution for the relative truth in 2025 for *common demersal* and *monkfish etc.* are positively correlated with each other (0.28), albeit weakly. This suggests that learning something about the *common demersal* group would tell you something about *monkfish etc.* Hence the mizer simulator gives some information regarding the *monkfish etc.* despite not actually predicting it.

See Appendix D for the other correlations between the groups.

4 Discussion

By treating the simulator outputs as coming from a population of simulators and modelling this population, we have presented in this paper a general way of combining ecosystem simulators to inform scientists and decision makers about the consequences of management strategies. Our model combines many different simulators, exploiting their strengths and discounting their weaknesses (Chandler 2013) to provide synthetic and comprehensive information to support decision making.

4.1 General model features

One of the difficulties in building an ensemble model with ecosystem simulators is that the simulator outputs are often done on different scales and are not directly comparable, for example StrathE2E models groups of species (e.g. pelagic, demersal) whereas mizer models major species individually. Our approach, unlike existing methods of combining simulators (e.g. Bayesian model averaging (Banner and Higgs 2017; Ianelli et al 2016)), allows us to combine outputs from these widely differing simulators. We achieve this by modelling what each simulator would predict for each of the groups of species we are interested in, whether it is explicitly modelled or not by the simulator. For example, in the case study, StrathE2E only models the total demersal species. Using information from the other simulators regarding the breakdown of demersal species and how the dynamics between species work, the ensemble model can say what StrathE2E would predict on a species level. In the case study, EwE and StrathE2E both implicitly predict groups of species. For EwE it is the sum of *poor cod and rays* and *other demersal* and for StrathE2E

it is the sums of all of the groups. As with the simulators that do not predict specific groups, we are able to infer what these simulators predict about implicit groups through correlations learned from other simulators. In this sense, the mizer model, which only predicts *common demersal* and *sole*, gives information about how StrathE2E divides its demersal species and therefore gives some information about other groups. Therefore, if we were interested in what would happen to the other demersals if we were to stop fishing, we should include all the simulators despite only two of them predicting it.

Simulators that are predictably wrong are more informative than those that are unpredictably wrong, even if the latter are less wrong in the absolute sense. In our framework, we distinguish between short-term and long-term individual discrepancies, which allows us to distinguish between predictably wrong simulators with small short-term individual discrepancies, $\boldsymbol{z}_i$, and unpredictably wrong simulators. Furthermore, we allow the short-term individual discrepancies to be different for each group, thus allowing a simulator to contribute to the ensemble model for groups that it is informative about and be ignored for groups that it is not. In the case study, mizer does not predict the dynamics of *sole* very well and so the simulator consensus, $\boldsymbol{\mu}$, only weakly follows the mizer predictions. On the other hand, mizer does a reasonable job of predicting the dynamics of *common demersal* and therefore it contributes more to the simulator consensus for this group. Thus the ensemble model exploits mizer's strengths, *common demersal*, and discounts its weaknesses, *sole*.

The ensemble model enables formal quantification of uncertainty. This uncertainty reflects a specific individual's updated beliefs having observed the simulators and the observation data (Robert 2007). The individual could be a scientist or a decision maker and could be informed by multiple experts (Albert et al 2012). Such

a framework could be used to help communicate uncertainty or enable decision-makers to directly quantify risks and therefore evaluate management trade-offs more rigorously (Harwood and Stokes 2003; Finkle 1990). The ensemble model takes account of uncertainty from each of the simulators, through parameter uncertainty and structural uncertainty, data uncertainty, through noisy and possibly indirect observations of the truth, and uncertainty in the ensemble model parameters.

As the simulators are describing the same system, we might expect the dynamics in the individual discrepancies to be similar. To reflect this, we allow the short-term individual discrepancies to come from some underlying distribution. Furthermore, in ecosystems simulators, the dynamics may be similar in direction but likely not in magnitude. To include this information in the case study, we split the short-term individual discrepancies, Λ_i, into correlations and magnitude (equation 3), allowing different levels of confidence for each. We used beta distributions for each of the off-diagonal elements of the correlation matrix and then conditioned on positive definiteness. This enabled us to learn about each element of the correlation matrix separately which is not possible in other formulations of the covariance matrix (Alvarez et al 2014). By acknowledging these features of simulators, we were able to better quantify the uncertainty.

It was also important to use informative priors as none of the simulators explicitly model *other demersal*. As there is no lower bound (on the log scale) for the values of the "best guess" of *other demersal*, we required some prior information about the distribution of the standard deviations, Π. This does suggest that the ensemble prediction is somewhat based on that of the priors for Λ_i. In practice, we suggest checking that your ensemble model predicts in a way that the decision maker believes before observing the truth, similar to the hypothetical data method of Kadane et al (1980). In the case study described here, we checked that the

dynamics of the biomasses prior to 1986 followed JLB's beliefs.

When building the ensemble model, how the species groups are decided depends on the question being asked. In the case study, we were interested in what would happen to demersal fish if we were to stop fishing, so we grouped the species into as few groups as possible. However, if we were interested in another question, for example if we had been interested in what would happen to commercial fish, we would divide the species into groups with commercial and non-commercial fish conditioned on species in each group being presented in exactly the same simulators. As the number of groups increases, the dimensions of the covariance matrices increases, so we advise that the number of groups be kept to a minimum as this would aid computation time and require less simulators and prior elicitation.

Using the ensemble model developed here, there is no need to identify the "best model" driven by the question being asked (Dickey-Collas et al 2014), but one should include all available simulators. Rather than developing many simulation models to answer different specific questions, the ensemble model can be designed to answer the question at hand thus reducing computational costs. Furthermore, as the ensemble model implicitly weights the simulators by their strengths and weaknesses, it is better for a simulator to be good at modelling one aspect of the ecosystem rather than being average at modelling a lot of things (Anderson et al 2017). Due to tractability it is not possible to explicitly show these weightings in the case study presented here, for an example of weightings in a more tractable example see Chandler (2013).

The nature of the different ecosystem simulators capturing different processes can limit the number of models available to run certain scenarios (e.g. in climate scenarios where some but not all the simulators contain links to temperature). If we were interested in one of the scenarios that a specific simulator was unable to run,

we should still include that simulators in the ensemble model as it gives information about how species interact with one another as well as the state of the ecosystem up until the current time. To include this simulator in the ensemble, we could learn about how it differs from the simulators that were able to run the specific scenario and increase a simulator's parameter uncertainty, Σ_i, as a function of time with in the future (Szuwalski and Thorson 2017).

4.2 Future work and extensions

Some ecosystem simulators are more similar than others, for example there are a number of size-based simulators in the marine literature (e.g. Blanchard et al 2009; Scott et al 2014) that are very similar, which may violate the exchangeability assumption made in Section 2. Additional hierarchy could be added to the ensemble model that would allow such simulators to have more similar discrepancies. In climate science, where the simulators are very similar to one another and phylogenetic trees show the development history of each simulator (Knutti et al 2013), Demetriou (2016) added additional hierarchy allowing closely related simulators to have similar discrepancies. They found that the major source of uncertainty was due to the shared discrepancy, and the results of the ensemble model were close to when all the simulators were assumed to be exchangeable.

In this paper, we have demonstrated the ideas and methods in cases where the quantities of interest are of fairly low dimension and have joint Gaussian distributions. However, with the increased efficiency of new statistical software and algorithms (see e.g. Girolami and Calderhead 2011), it is possible to address larger problems involving more general distributions.

The framework presented here is not exclusive to ecosystem simulators in fisheries, but can be used to combine any mechanistic simulators in many areas of

ecology (e.g. Individual-based models, Railsback and Grimm 2012) or even other areas of research such as systems biology (Kuepfer et al 2007) and epidemiology (Lessler et al 2016).

4.3 Conclusion

This work allows for a synthesis of many modelling studies that have been and are being conducted in such a way that we can obtain more holistic knowledge over a wide scope of complex ecological systems. It also allows for including a formal quantitative understanding of uncertainties and knowledge gaps. This enables us to make comprehensive model projections that take into account all that we have learnt from the simulators collectively.

Acknowledgments

The work was supported by the Natural Environment Research Council and Department for Environment, Food and Rural Affairs [grant number NE/L003279/1, Marine Ecosystems Research Programme]. The authors would like to thank Tom Webb, Remi Vergnon, Yuri Artioli, Sévrine Saillery, Paul Somerfield, Melanie Austen, Nicola Beaumont and Stefanie Broszeit for participating in early elicitation exercises. We thank Tony Pitcher and two anonymous reviewers for comments on an earlier version of the paper.

Author contribution

MAS, PGB and JLB conceived the ideas and designed the methodology; JLB extracted the data for the main case study; MAS, MRH, SM, DS, AGR, RBT, JJH

and NS ran the simulators for the case study; MAS implemented the methodology; MAS and PGB analysed the data; MAS and PGB led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Albert I, Donnet S, Guihenneuc-Jouyaux C, Low-Choy S, Mengersen K, Rousseau J (2012) Combining Expert Opinions in Prior Elicitation. Bayesian Analysis 7(3):503–532

Alhussain ZA, Oakley JE (2017) Eliciting judgements about uncertain population means and variances. arXiv:170200978 URL https://arxiv.org/abs/1702.00978

Alvarez I, Niemi J, Simpson M (2014) Bayesian inference for a covariance matrix. arXiv:14084050 URL https://arxiv.org/abs/1408.4050

Anderson SC, Cooper AB, Jensen OP, Minto C, Thorson JT, Walsh JC, Afflerbach J, Dickey-Collas M, Kleisner KM, Longo C, Osio GC, Ovando D, Mosqueira I, Rosenberg AA, Selig ER (2017) Improving estimates of population status and trend with superensemble models. Fish and Fisheries DOI 10.1111/faf.12200

Banner KM, Higgs MD (2017) Considerations for Assessing Model Averaging of Regression Coefficients. Ecological Applications 27(1):78–93, DOI 10.1002/eap.1419

Berger JO (1985) Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer Series in Statistics, Springer-Verlag

Blanchard JL, Jennings S, Law R, Castle MD, McCloghrie P, Rochet MJ, Benoît

E (2009) How does abundance scale with body size coupled size-structured food webs? Journal of Animal Ecology 78(270-280)

Blanchard JL, Andersen KH, Scott F, Hintzen NT, Piet G, Jennings S (2014) Evaluating targets and trade-offs among fisheries and conservation objectives using multispecies size spectrum model. Journal of Applied Ecology 51(3):612–662, DOI 10.1111/1365-2664.12238

Chandler RE (2013) Exploiting strength, discounting weakness: combining information from multiple climate simulators. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371(1991), DOI DOI: 10.1098/rsta.2012.0388

Demetriou D (2016) A Bayesian approach to the interpretation of climate model ensembles. PhD thesis, University College London

Dickey-Collas M, Payne MR, Trenkel VM, Nash RDM (2014) Hazard warning: model misuse ahead. ICES Journal of Marine Science: Journal du Conseil 72(8):2300–2306

Finkle AM (1990) Confronting uncertainty in risk management: A guide for decision-makers: a report. Tech. rep., Centre for Risk Management, Resourses for the Future

Fung T, Farnsworth KD, Reid DG, Rossberg AG (2012) Recent data suggests no further recovery in North Sea Large Fish Indicator. ICES Journal of Marine Science 69:235–239, DOI 10.1093/icesjms/fsr206

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian Data Analysis, 3rd edn. Chapman and Hall

578 Gelman A, Lee D, Guo J (2015) Stan: A probabilistic programming language.
579 Journal of Educational and Behavioural Statistics 40:530–543

580 Girolami M, Calderhead B (2011) Riemann manifold Langevin and Hamiltonian
581 Monte Carlo methods. Journal of Royal Statistical Society B 73:1–37, DOI
582 10.1111/j.1467-9868.2010.00765.x

583 Harwood J, Stokes K (2003) Coping with uncertainty in ecological advice: lessons
584 from fisheries. Trends in Ecology and Evolution 18(12):617–622

585 Heath MR (2012) Ecosystem limits to food web fluxes and fisheries yields in the
586 north sea simulated with an end-to-end food web model. Progress in Oceanogra-
587 phy 102:42 – 66, DOI 10.1016/j.pocean.2012.03.004

588 Heath MR, Speirs DC, Steele JH (2014) Understanding patterns and processes in
589 models of trophic cascades. Ecology Letters 17:101–114, DOI 10.1111/ele.12200

590 Hoffman MD, Gelman A (2014) The No-U-Turn Sampler: Adaptively Setting Path
591 Lengths in Hamilitonian Monte Carlo. Journal of Machine Learning Research
592 15:1593–1623

593 Hyder K, Rossberg AG, Allen JI, Austen MC, Barciela RM, Bannister HJ, Blackwell
594 PG, Blanchard JL, Burrows MT, Defriez E, Dorrington T, Edwards KP, Garcia-
595 Carreras B, Heath MR, Hembury DJ, Heymans JJ, Holt J, Houle JE, Jennings S,
596 Mackinson S, Malcolm SJ, McPike R, Mee L, Mills DK, Montgomery C, Pearson
597 D, Pinnegar JK, Pollicino M, Popova EE, Rae L, Rogers SI, Speirs D, Spence
598 MA, Thorpe R, Turner RK, van der Molen J, Yool A, Paterson DM (2015)
599 Making modelling count - increasing the contribution of shelf-seas community
600 and ecosystem models to policy development and management. Marine Policy
601 61:291 – 302, DOI 10.1016/j.marpol.2015.07.015

602 Ianelli J, Holsman KK, Punt AE, Aydin K (2016) Multi-model inference for in-
603 corporating trophic and climate uncertainty into stock assessments. Deep Sea
604 Research Part II: Topical Studies in Oceanography 134:379–389

605 ICES Database of Trawl Surveys (DATRAS) (2015) International Bottom Trawl
606 Survey (IBTS) data 1985-2014. URL http://datras.ices.dk

607 Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends
608 in Ecology & Evolution 19(2):101–108, DOI 10.1016/j.tree.2003.10.013

609 Kadane J, Dickey J, Winkler J, Smith W, Peters S (1980) Interactive elicitation of
610 opinion for a normal linear-model. Journal of American Statistical Association
611 75(372):845–854

612 Knutti R (2010) The end of model democracy? Climate Change 102:395–404

613 Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: Generation
614 CMIP5 and how we got there. Geophysical Research Letters 40(6):1194–1199,
615 DOI 10.1002/grl.50256

616 Kuepfer L, Peter M, Sauer U, Stelling J (2007) Ensemble modeling for anal-
617 ysis of cell signaling dynamics. Nature Biotechnology 25(9):1001–1006, DOI
618 10.1038/nbt1330

619 Leith NA, Chandler RE (2010) A framework for interpreting climate model outputs.
620 Journal of the Royal Statistical Society: Series C (Applied Statistics) 59(2):279–
621 296, DOI 10.1111/j.1467-9876.2009.00694.x

622 Lessler J, Azman AS, Grabowski MK, Salje H, Rodriguez-Barraquer I (2016) Trends
623 in the mechanistic and dynamic modeling of infectious diseases. Current Epidemi-
624 ology Reports 3(3):212–222, DOI 10.1007/s40471-016-0078-4

Li H, Wu J (2006) Uncertainty analysis in ecological studies. In: Wu J, Jones KB, Li H, Loucks OL (eds) Scaling and Uncertainty Analysis in Ecology: Methods and Applications, 43-64, Springer, pp 43–64

Lynam CP, Mackinson S (2015) How will fisheries management measures contribute towards the attainment of good environmental status for the North Sea ecosystem? Global Ecology and Conservation 4(0):160–175, DOI 10.1016/j.gecco.2015.06.005

Mackinson S, Platts M, Garcia C, Lynam CP (2018) Evaluating the fishery and ecological consequences of the proposed North Sea multi-annual plan. PLOS ONE 13(1):e0190015. DOI 10.1371/journal.pone.0190015

Morris DJ, Speirs DC, Cameron AI, Heath MR (2014) Global sensitivity analysis of an end-to-end marine ecosystem model of the North Sea: Factors affecting the biomass of fish and benthos. Ecological Modelling 273:251–263

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. John Wiley and Sons

Payne MR, Barange M, Cheung WWL, MacKenzie BR, Batchelder HP, Cormon X, Eddy TD, Fernandes JA, Hollowed AB, Jones MC, Link JS, Neubauer P, Ortiz I, Queirós AM, Paula J (2015) Uncertainties in projecting climate-change impacts in marine ecosystems. ICES Journal of Marine Science: Journal du Conseil DOI 10.1093/icesjms/fsv231

Railsback SF, Grimm V (2012) Agent-based and individual-based modeling a practical introduction. Princeton University Press

Robert CP (2007) The Bayesian Choice, 2nd edn. Springer, New York

649 Rougier J (2016) Ensemble Averaging and Mean Squared Error. Journal of Climate
650 29(24):8865–8870, DOI 10.1175/JCLI-D-16-0012.1

651 Rougier J, Goldstein M, House L (2013) Second-Order Exchangeability Anal-
652 ysis for Multimodel Ensembles. Journal of American Statistical Association
653 108(503):852–863

654 Scott F, Blanchard JL, Andersen KH (2014) mizer: an R package for multispecies,
655 trait-based and community size spectrum ecological modelling. Methods in Ecol-
656 ogy and Evolution 5(10):1121–1125, DOI 10.1111/2041-210X.12256

657 Speirs D, Guirey E, Gurney W, Heath M (2010) A length-structured partial ecosys-
658 tem model for cod in the north sea. Fisheries Research 106(3):474 – 494, DOI
659 10.1016/j.fishres.2010.09.023

660 Spence MA, Blackwell PG, Blanchard JL (2016) Parameter uncertainty of a dy-
661 namic multi-species size spectrum model. Canadian Journal of Fisheries and
662 Aquatic Sciences 73(4):589–597

663 Szuwalski CS, Thorson JT (2017) Global fishery dynamics are poorly predicted by
664 classical models. Fish and Fisheries DOI 10.1111/faf.12226

665 Tebaldi C, Sansó B (2009) Joint projections of temperature and precipita-
666 tion change from multiple climate models: a hierarchical Bayesian approach.
667 Journal of Royal Statistics Society A 172(1):83–106, DOI 10.1111/j.1467-
668 985X.2008.00545.x

669 Thorpe RB, Le Quesne WJF, Luxford F, Collie JS, Jennings S (2015) Evaluation
670 and management implications of uncertainty in a multi-species size-structured
671 model of population and community responses to fishing. Methods in Ecology
672 and Evolution 6(1):49–58

673 Tittensor DP, Eddy TD, Lotze HK, Galbraith ED, Cheung W, Barange M, Blan-
674 chard JL, Bopp L, Bryndum-Buchholz A, Büchner M, Bulman C, Carozza DA,
675 Christensen V, Coll M, Dunne JP, Fernandes JA, Fulton EA, Hobday AJ, Huber
676 V, Jennings S, Jones M, Lehodey P, Link JS, Mackinson S, Maury O, Niiranen S,
677 Oliveros-Ramos R, Roy T, Schewe J, Shin YJ, Stock CA, Underwood P, Volkholz
678 J, Watson JR, Walker ND (2017) A protocol for the intercomparison of marine
679 fishery and ecosystem models: Fish-MIP v1.0. Geoscientific Model Development
680 Discussions 2017:1–39, DOI 10.5194/gmd-2017-209

681 Walker ND, Maxwell DL, Le Quesne WJF, Jennings S (2017) Estimating efficiency
682 of survey and commercial trawl gears from comparisons of catch-ratios. ICES
683 Journal of Marine Science 74(5):1448–1457, DOI 10.1093/icesjms/fsw250

684 Williams PJ, Hooten MB (2016) Combining statistical inference and decisions in
685 ecology. Ecological Applications 26(6):1930–1942, DOI 10.1890/15-1593.1

Tables

Table 1: A summary of the variables in the ensemble model. The ensemble model is run for $t = 1 \dots T$.

Variable	Dimension	Times	Description	Relationship
$\boldsymbol{y}^{(t)}$	n	$t = 1 \dots T$	The truth	$\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)} + \boldsymbol{\epsilon}_{\Lambda,t}$
$\boldsymbol{w}^{(t)}$	n_y	$t = 1 \dots T$	Possibly incomplete version of the truth	$\boldsymbol{w}^{(t)} = f_y(\boldsymbol{y}^{(t)})$
$\hat{\boldsymbol{w}}^{(t)}$	n_y	$t \in S_0$	Noisy observation of $\boldsymbol{w}^{(t)}$	$\hat{\boldsymbol{w}}^{(t)} \sim p(\hat{\boldsymbol{w}}^{(t)} \vert \boldsymbol{w}^{(t)})$
$\boldsymbol{\delta}$	n	NA	Long-term shared discrepancy	
$\boldsymbol{\eta}^{(t)}$	n	$t = 1 \dots T$	Short-term shared discrepancy	$\boldsymbol{\eta}^{(t)} = R_\eta \boldsymbol{\eta}^{(t-1)} + \boldsymbol{\epsilon}_{\eta,t}$
$\boldsymbol{\mu}^{(t)}$	n	$t = 1 \dots T$	Simulator concensus	$\boldsymbol{\mu}^{(t)} = \boldsymbol{y}^{(t)} + \boldsymbol{\delta} + \boldsymbol{\eta}^{(t)}$
$\boldsymbol{\gamma}_i$	n	NA	Simulator i's long-term individual discrepancy	
$\boldsymbol{z}_i^{(t)}$	n	$t = 1 \dots T$	Simulator i's short-term individual discrepancy	$\boldsymbol{z}_i^{(t)} = R_i \boldsymbol{z}_i^{(t-1)} + \boldsymbol{\epsilon}_{z,t,i}$
$\boldsymbol{x}_i^{(t)}$	n	$t = 1 \dots T$	Simulator i's best guess	$\boldsymbol{x}_i^{(t)} = \boldsymbol{\mu}^{(t)} + \boldsymbol{\gamma}_i + \boldsymbol{z}_i^{(t)}$
$\boldsymbol{u}_i^{(t)}$	n_i	$t = 1 \dots T$	Simulator i's incomplete version of $\boldsymbol{x}_i^{(t)}$	$\boldsymbol{u}_i^{(t)} = f_i(\boldsymbol{x}_{\boldsymbol{i}}^{(t)})$
$\hat{\boldsymbol{u}}_i^{(t)}$	n_i	$t \in S_i$	The expectation of simulator i's output $\boldsymbol{u}_i^{(t)}$	$\boldsymbol{u}_i^{(t)} = \hat{\boldsymbol{u}}_i^{(t)} + \boldsymbol{\epsilon}_{u_i}$

Table 2: A summary of the simulators, their outputs used in the case study, the simulator-specific function, $\boldsymbol{u}_i^{(t)} = f_i \boldsymbol{x}_i^{(t)} = M_i 10^{\boldsymbol{x}_i^{(t)}}$ and a reference to where the parameter uncertainty, Σ_i, was calculated.

Simulator	Description	Outputs	M_i	Reference for Σ_i
EcoPath with EcoSim (EwE)	Total biomass is modelled at the species level	1) *Common demersal* 2) *Sole* 3) *Monkfish etc.* 4) Sum of *Poor cod and Rays* and *Other demersal fish* for $t = 1991 - 2023$.	$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$.	Mackinson et al (2018)
mizer	Total weight is modelled in weight classes by species	1) *Common demersal* 2) *Sole* for $t = 1968 - 2100$.	$M_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$.	Spence et al (2016)

FishSUMs	Abundance in length classes is modelled by species	1) *Common demersal* for $t = 1990 - 2098$.	$M_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}$.	This study, see Appendix B
StrathE2E	Biomass is modelled for different functional groups	1) Sum of *Common demersal*, *Sole*, *Monkfish etc.*, *Poor cod and Rays* and *Other demersal fish* for $t = 1983 - 2050$.	$M_4 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$.	This study, see Appendix B
LeMans	Abundance in length classes is modelled by species	1) *Common demersal* 2) *Sole* 3) *Monkfish etc.* 4) *Poor cod and Rays* for $t = 2000 - 2099$	$M_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$.	Thorpe et al (2015)

Figures

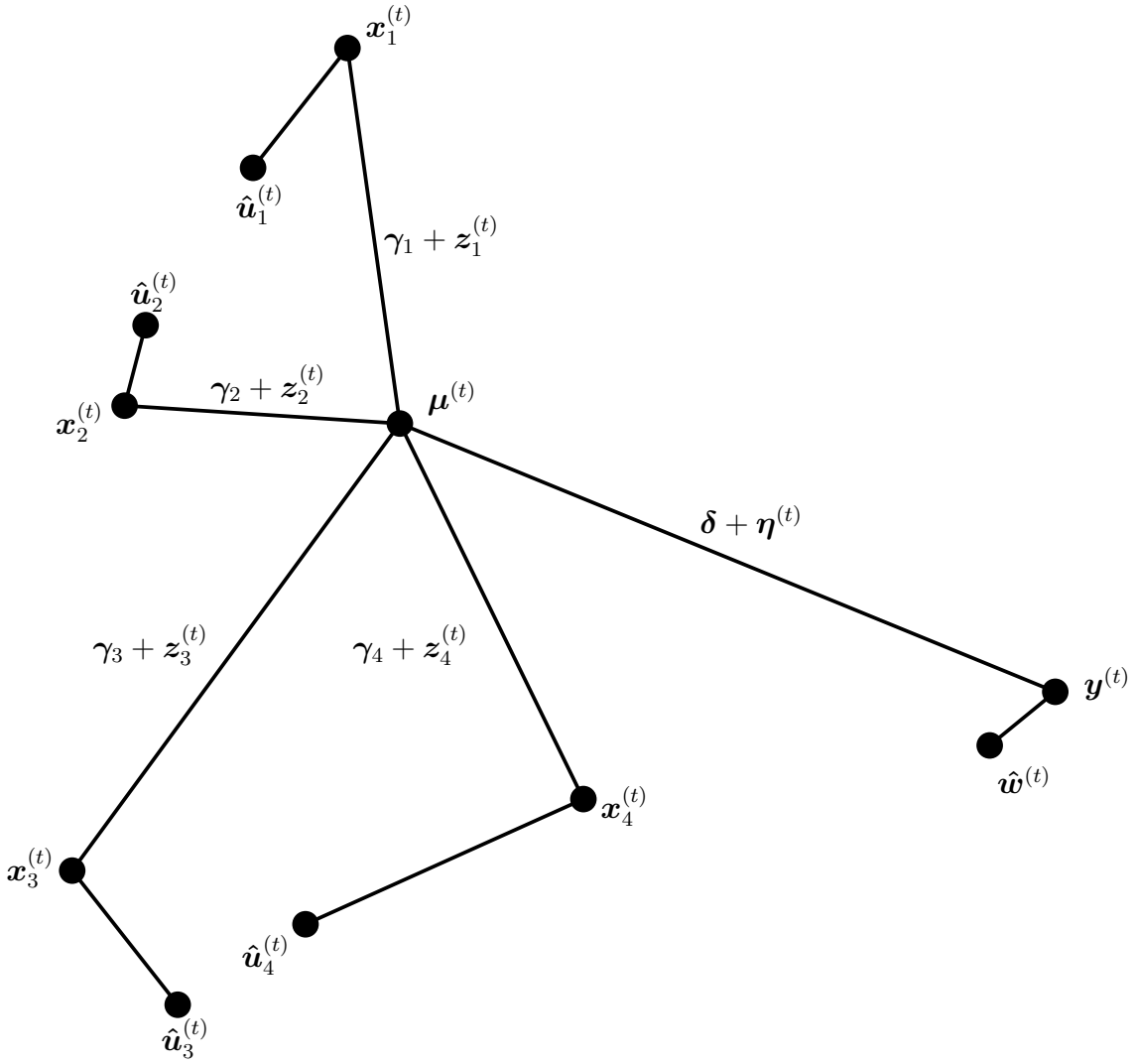

Figure 1: A schematic that shows an example of the ensemble model at time t. In this example we have four simulators that are all able to predict the elements of $\boldsymbol{y}^{(t)}$. Each simulator's 'best guess', $\boldsymbol{x}_i^{(t)}$, is observed with parameter uncertainty where $\hat{\boldsymbol{u}}_i^{(t)}$ is the expected output of the ith simulator (see Section 2.1). The difference between the ith simulator's 'best guess', $\boldsymbol{x}_i^{(t)}$, and the simulator consensus, $\boldsymbol{\mu}^{(t)}$, is known as simulator i's individual discrepancy and is split between its long-term, $\boldsymbol{\gamma}_i$, and short-term, $\boldsymbol{z}_i^{(t)}$, individual discrepancy (see Section 2.2). The difference between the truth, $\boldsymbol{y}^{(t)}$ and the

simulator consensus, $\boldsymbol{\mu}^{(t)}$, is known as the shared discrepancy and is divided into long-term, $\boldsymbol{\delta}$, and short-term, $\boldsymbol{\eta}^{(t)}$, shared discrepancy (see Section 2.3). In addition, we do not directly observe the truth but we do observe a noisy version of it, $\hat{\boldsymbol{w}}^{(t)}$ (see Section 2.4).

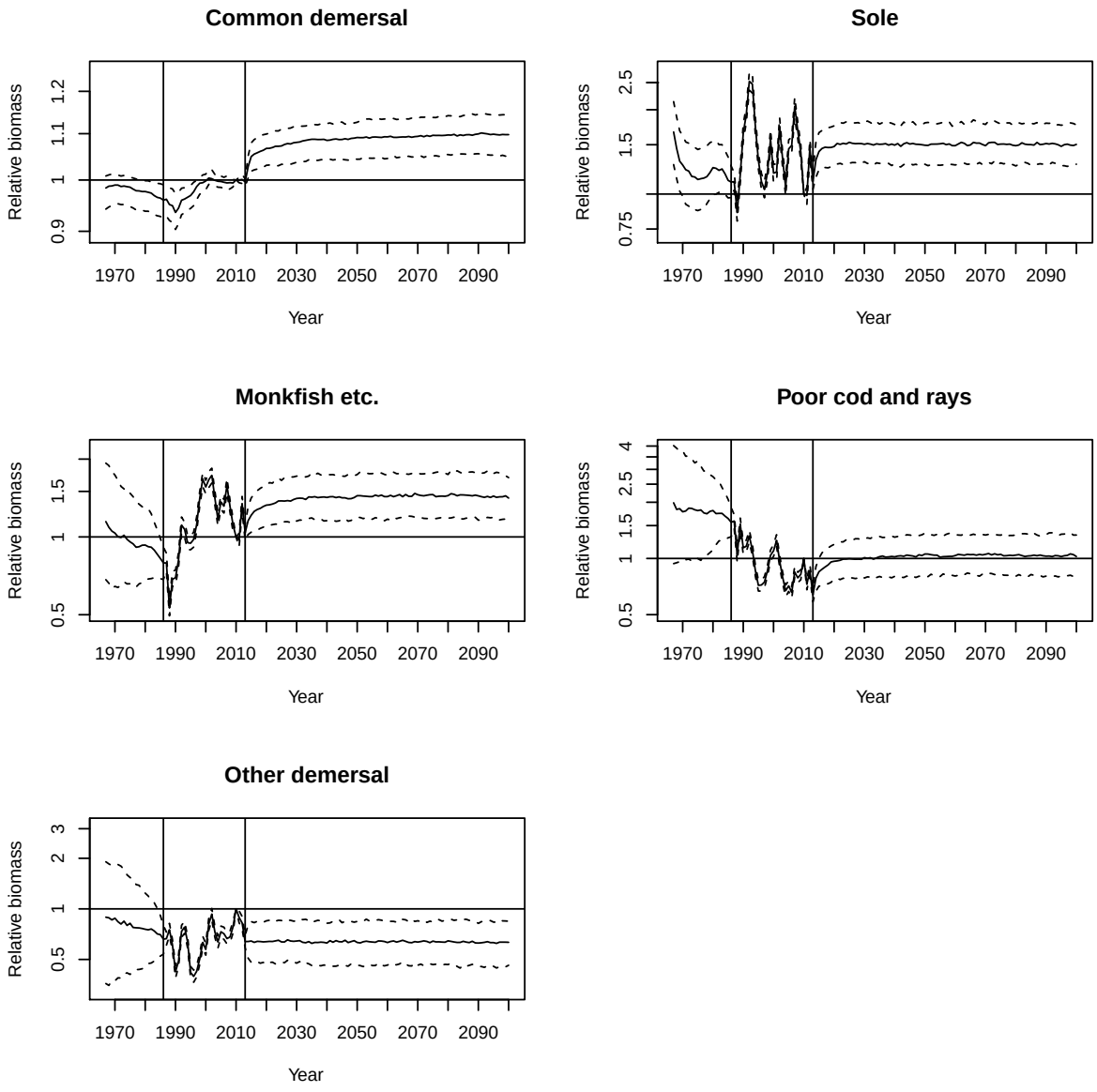

Figure 2: Estimates of the log biomass of each group of species relative to 2010. The solid line is the median and the dotted lines are the upper and lower quartiles. The first vertical line is at 1986, the year that we first have data, and the second line is in 2013, the simulated cessation of fishing.

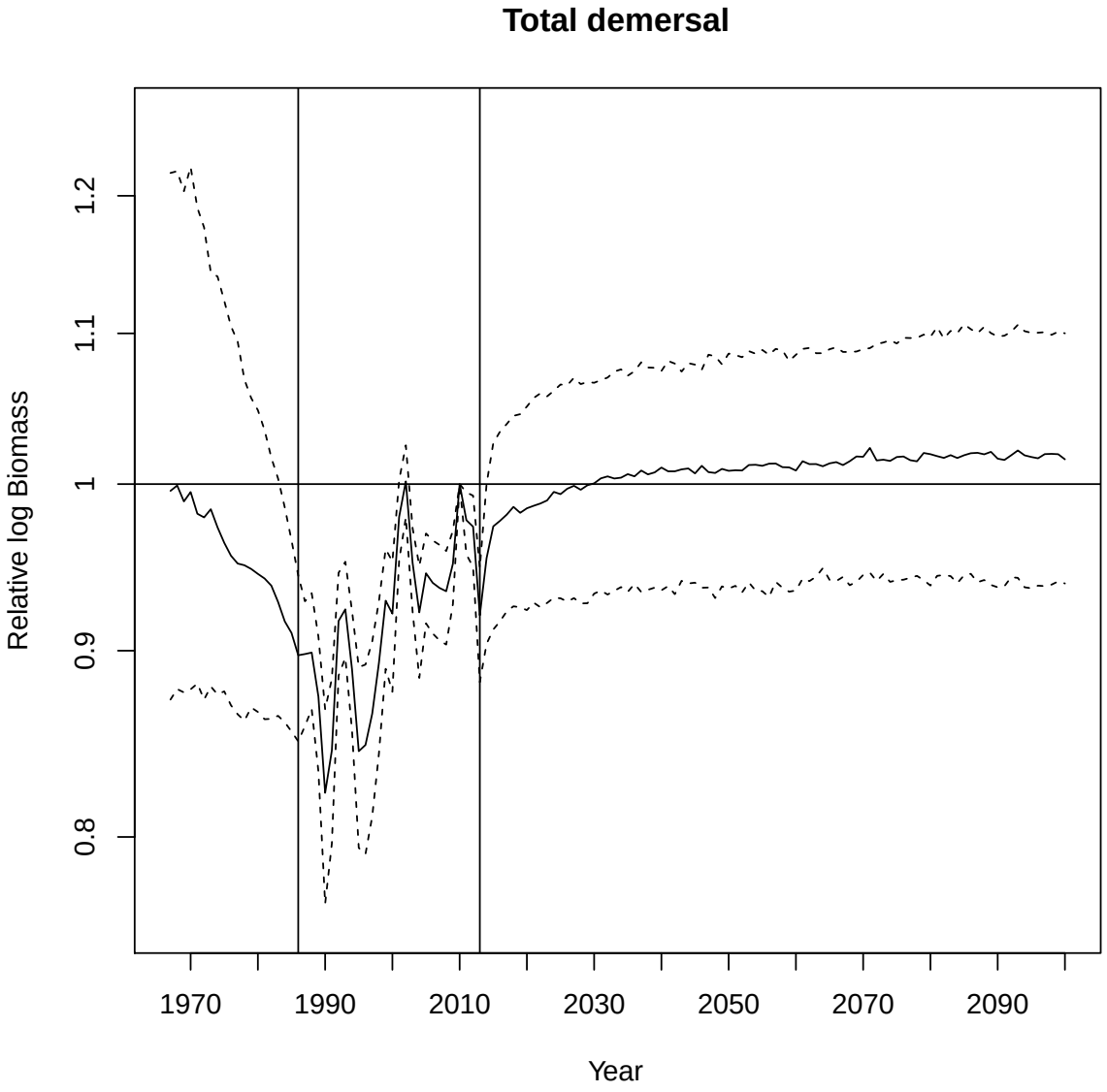

Figure 3: The total biomass of demersal species as predicted by the models relative to 2010.

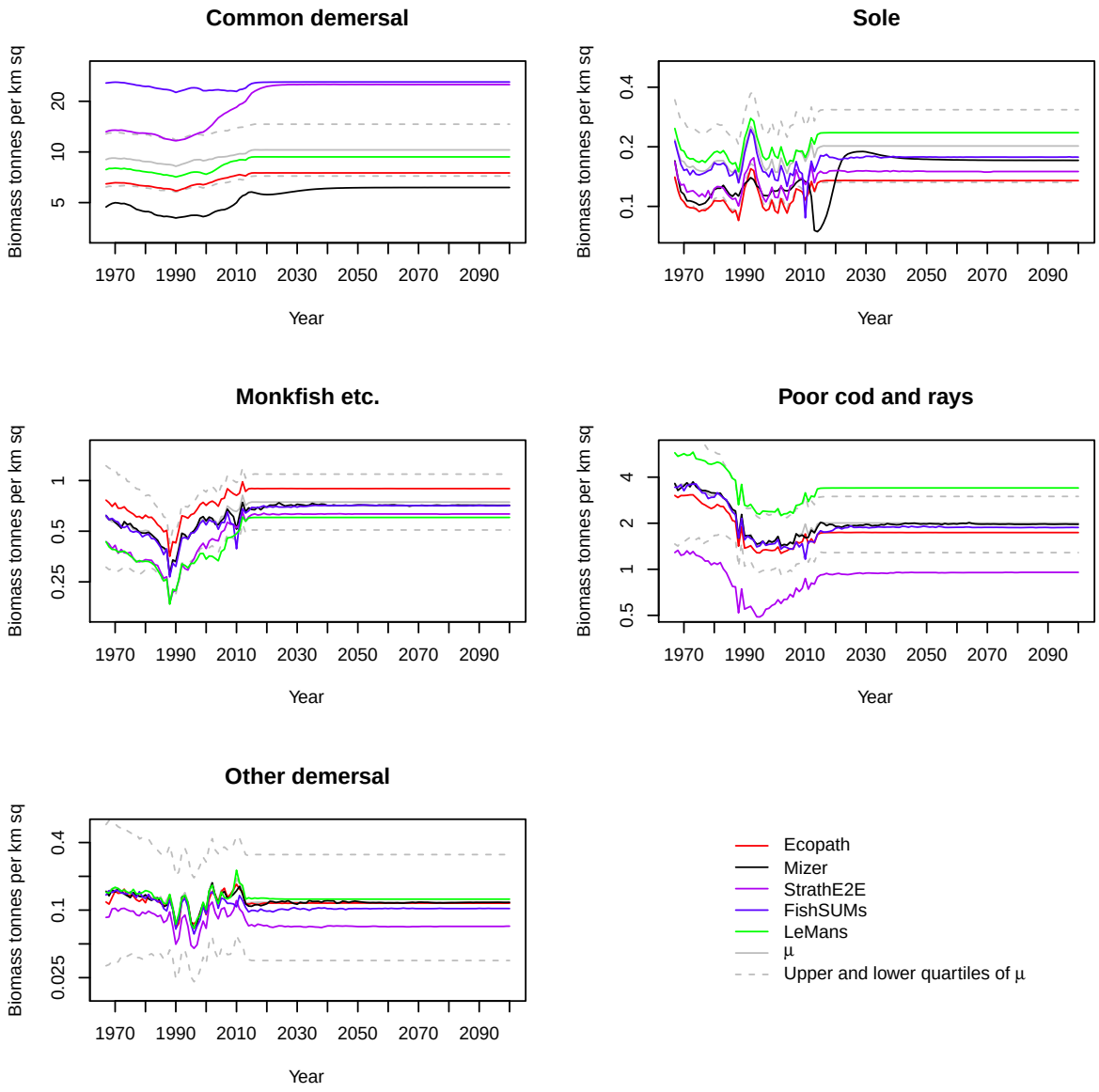

Figure 4: The median best guess for the simulators ($\boldsymbol{x}_i$) for mizer (black), FishSUMs (purple), LeMans (green), EwE (red) and StrathE2E (pink) and the median simulator consensus ($\boldsymbol{\mu}$) and its quartiles in solid grey and dotted grey respectively.